15 September 2021

Elham Tabassi and Mark Przybocki
National Institute of Standards and Technology
MS 20899, 100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI Risk Management Framework

Via email to AIframework@nist.gov

Dear Ms. Tabassi and Mr. Przybocki,

Thank you for the invitation to submit comments in response to the National Institute of Standards and Technology (NIST) Request for Information (RFI) on the NIST AI Risk Management Framework (AI RMF or Framework). NIST requested comments be sent to AIframework@nist.gov or www.regulations.gov. We offer the following submission for your consideration.

We focus on three broad categories of risks: to *democracy and security*, to *human rights and well-being*, and of *global catastrophes*. Although many real-world examples of risks may fit into more than one of those categories, each category also has important analytical distinctions and is independently important for ensuring that the future development of AI systems remains safe and commensurate with human priorities.

While prior work has argued for treating each type of risk seriously and urgently, we emphasize that these risks—however unlikely or difficult to imagine today—are likely to reverberate and exacerbate each other unless we properly address and mitigate them. Put differently, we cannot exhaustively prepare for any of these risks unless due attention is paid to each of them. This entails active monitoring and proactive mechanisms to prevent their manifestation and mutual effects. Consequently, the gap we aim to fill with this submission to NIST is the identification of policy strategies, institutional mechanisms, and technical interventions that speak to the intersection of these risks, with emphasis on themes that cut across the particular dangers or warnings articulated by AI theorists, computer scientists, policymakers, and stakeholder advocates.

Our key general topics and recommendations include:

- Keep focusing on and delineate the meaning of societal-scale issues, to include: risks to democracy and security; risks to human rights and wellbeing; and global catastrophic risks.
    - We appreciate that NIST has dedicated substantial attention to societal-scale issues in the AI RMF RFI, in addition to individual and group risks.
    - We recommend that the meaning of societal scale issues be expanded to

include: risks to democracy and security, such as polarization, extremism, mis- and disinformation, and social manipulation; risks to human rights and wellbeing including equity, environmental, and public health risks; and global catastrophic risks, including risks to large numbers of people caused by AI accidents, misuse, or unintended impacts in both the near- and long-term.

- Risk assessment approaches focused on intended use cases have important limitations.
    - Consideration of intended AI use-cases is valuable and necessary, but not sufficient, for identification and assessment of important AI risks.
    - We appreciate that NIST goes beyond focusing on intended use cases in the RMF RFI.
    - We recommend that the RMF include clear, usable guidance on identifying and assessing risks of AI, yielding risk management strategies that would be robust despite high uncertainty about future potential uses and misuses beyond the AI designers' originally intended/planned uses.
- The nascent but growing field of AI safety is providing insights about AI risks and risk management.
    - While much of the work in the field of AI safety is at an early stage, it has already yielded some general principles and tools that we expect could be useful to NIST stakeholders.
    - We recommend that the NIST Framework consider the nascent but growing field of AI safety in informing its deliberations.
- NIST should continue to maintain awareness of progress in AI safety and other key fields, and update corresponding components of the RMF as needed.
    - The AI field has changed significantly over the last five years, and is likely to continue to change, perhaps even more dramatically.
    - We recommend that NIST maintain close relationships with researchers in key fields (including AI safety, security, and capabilities) to follow shifts across these fields and potential impact on the RMF, and that NIST update corresponding components of the Framework as needed.
- Coordination of standards for risk identification and mitigation, to the extent possible.
    - We recommend that NIST be explicit about how and where the RMF will incorporate and coordinate with existing and future AI standards development and risk assessment.

In the following sections, we first expand on our above comments related to key cross-cutting general RMF RFI topics, with a focus on the aforementioned categories of risk (to democracy and security, to human rights and well-being, and global catastrophic risks). We describe each type of risk in detail, outline various subsets and examples, and highlight existing technical and policy work that speaks to them. We then provide separate comments on specific RFI topics. Our recommendations in response to the specific RFI topics include the following:

- We recommend that the RMF provide guidance on risk identification, assessment, and prioritization processes to include risks that could have high consequences for society but may seem to AI designers to be outside the typical scope of consideration for their

organization, such as events that would be novel or low-probability events, or systemic risks, or expected to be outside their typical time horizon. (Recommendation for RFI Topic 1)

- We recommend that NIST consult with a diverse set of stakeholders, including risk-sensitive groups, for input such as on definitions of key terms to better understand how the terms have been used differently by various stakeholders. (Recommendation for RFI Topic 2)
- We recommend that NIST consider "assessment of generality" (i.e., assessment of the breadth of AI applicability/adaptability) as another important characteristic affecting trustworthiness of an AI, or perhaps as a factor affecting one or more of the AI trustworthiness characteristics NIST has already outlined. (Recommendation for RFI Topic 2)
- We recommend that NIST consider including principles of sustainability and inclusivity. We also recommend that NIST clarify two items in the RMF RFI regarding NIST's use of the terms "characteristics" and "principles": 1. That the difference between principles and characteristics is made more clear, and 2. Where the RFI states that "These characteristics and principles are generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services", we recommend you clarify to what extent NIST meant "considered by the public", or "considered by experts", or both. (Recommendation for RFI Topic 3)
- We recommend that NIST consider having the RMF include guidance to have risk identification processes performed by a team that is diverse, multidisciplinary, representing multiple departments of the organization, as well as including a correspondingly diverse set of stakeholders from outside the organization. (Recommendation for RFI Topic 5)
- We recommend that the RMF include standardized templates for reporting information on AI risk factors and incidents, that AI developers could adopt voluntarily. (Recommendation for RFI Topic 5)
- We recommend that NIST consider adding usability as an attribute of the AI RMF. (Recommendation for RFI Topic 9)
- We recommend that NIST consider clarifying its planned procedures for making RMF updates (how often, under what conditions, decision criteria), and how it aims to balance flexibility with standard-setting authority. (Recommendation for RFI Topic 10)
- We strongly recommend that the Framework include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. These should include guidance for determining who should be responsible for implementing the Framework within each organization, ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems, support for incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers, among other mechanisms. We also recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety. (Recommendation for RFI Topic 12)

Thank you again for the opportunity to comment on this RFI. If you need additional information or would like to discuss further, please let us know. In any case, we look forward to further engagement with NIST as you proceed on the RMF development process.

Our best,

Anthony Barrett, Ph.D., PMP
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Thomas Krendl Gilbert, Ph.D.
Research Affiliate
Center for Human-Compatible AI, UC Berkeley

Caroline Jeanmaire
Director of Strategic Research and Partnerships
Center for Human-Compatible AI, UC Berkeley

Jessica Newman
Program Lead
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Brandie Nonnecke, Ph.D.
Director
CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley

Ifejesu Ogunleye
Graduate Researcher
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Outline

General RMF RFI Topics 6
- Keep focusing on and delineate the meaning of societal-scale issues, to include: risks to democracy and security; risks to human rights and wellbeing; and global catastrophic risks. 6
  - Comments 6
  - Recommendations 10
- Risk assessment approaches focused on intended use cases have important limitations. 10
  - Comments 10
  - Recommendations 12
- The nascent but growing field of AI safety is providing insights about AI risks and risk management. 12

Comments 12
Recommendations 14
NIST should continue to maintain awareness of progress in AI safety and other key fields, and update corresponding components of the RMF as needed. 14
Comments 14
Recommendations 14
Coordination of standards for risk identification and mitigation, to the extent possible. 14
Comments 14
Recommendations 15
Specific RMF RFI Topics 15
RFI Topic 1 (The Greatest Challenges in Improving how AI Actors Manage AI-related Risks) 15
Comments 15
Recommendations 16
RFI Topic 2 (Characteristics of AI Trustworthiness) 17
Comments 17
Recommendations 18
RFI Topic 3 (AI Principles) 18
Comments 18
Recommendations 19
RFI Topic 4 (AI Risks in Organizations' Enterprise Risk Management) 20
Comments 20
RFI Topic 5 (Standards, Frameworks, Models, Methodologies, Tools, Guidelines and Best Practices) 20
Comments 20
Recommendations 22
RFI Topic 7 (Alignment with Other Efforts) 22
Comments 22
RFI Topic 8 (Inclusiveness) 23
Comments 23
RFI Topic 9 (Attributes for RMF) 23
Comments 23
Recommendations 23
RFI Topic 10 (Structuring the Framework) 24
Comments 24
Recommendations 25
RFI Topic 12 (Governance) 25
Comments 25
Recommendations 27

General RMF RFI Topics

Keep focusing on and delineate the meaning of societal-scale issues, to include: risks to democracy and security; risks to human rights and wellbeing; and global catastrophic risks.

Comments

We appreciate that NIST has dedicated substantial attention to societal-scale issues in the AI RMF RFI, in addition to individual and group risks. We recommend that the focus of impacts on society remain and for the meaning of societal-scale issues to be expanded to include:

1. Risks to democracy and security such as polarization, extremism, disinformation, and social manipulation;
2. Risks to human rights and wellbeing including equity, environmental, and public health risks; and
3. Global catastrophic risks including risks to large numbers of people caused by AI accidents, misuse, or unintended impacts in both the near- and long-term.

These categories are not mutually exclusive, and other categories also could be worth including.

Risks to Democracy and Security

Societal risks include that personalized disinformation (enabled by AI) on social media (e.g., through Twitter bots, synthesis of massive datasets from Facebook, deepfake videos) can sway elections (Brkan 2019) and incite genocide (Mozur 2018). AI-enabled automated surveillance systems could suppress dissent, and hackers can use AI to augment their capability for cyberattacks, including on critical infrastructure (Brundage et al. 2018).

Risks to Human Rights and Wellbeing

In addition to risks to democracy from AI-enabled disinformation, we have also seen throughout the COVID-19 pandemic the role of mis- and disinformation on public health outcomes, which is a major component of human rights and wellbeing.

The 2021 National Defense Authorization Act (NDAA) authorizes the Secretary of Commerce to establish a National AI Advisory Committee, including a Subcommittee on AI and Law Enforcement to provide guidance on issues such as privacy rights, civil rights and civil liberties, and disability rights implicated in the use of AI in law enforcement. We believe this work should be expanded beyond law enforcement and, at minimum, inform NIST's development of the AI Risk Management Framework.

As part of its focus on societal-scale risks, the RMF should include a rights-based approach in

the identification and mitigation of risks. This would include context-dependent harms pertaining to human rights violations, in addition to the scale-dependent harms of catastrophic or civilizational risk. The 2021 NDAA (specifically HR 6395, Division E, Section 5301(c)), authorizes NIST to provide definitions for AI trustworthiness concepts such as “privacy”, “fairness”, and “bias.” Several of these relate to recognized human rights in international law. For example, privacy is a protection granted under Article 12 of the Universal Declaration of Human Rights (UDHR), and fairness and bias are related to rights of “non-discrimination” protected under Article 2 of the UDHR and Article 26 in the International Covenant on Civil and Political Rights (ICCPR). These definitions have been codified over decades through the human rights legal framework (e.g., in charters and national laws and regulations).

The operationalization of “AI trustworthiness” concepts (e.g., non-discrimination) can be informed by the human rights legal framework through its operationalization of these concepts across international contexts and domain application areas. For example, the principle of “non-discrimination” can be understood through its interpretation in human rights law (e.g., its codification in relevant charters, case law, and regulations). By analyzing how responsible AI principles have been interpreted through the human rights legal framework, NIST can use definitions that have reached fairly widespread consensus over decades of negotiation. Relevant work on human rights in AI includes Nonnecke and Dawson (forthcoming), Latonero (2018), Donahoe and Metzger (2019), Mantelero and Esposito (2021), and Bradley et al. (2021).

As another area related to wellbeing, environmental risks include significant energy costs of deep learning and resource extraction for computational hardware and chips. The trend toward ever-larger language models exacerbates these risks (Bender et al. 2021).

Global Catastrophic Risks

AI systems could pose risks of catastrophe from malicious or unintentional misuse, accidents, or other failures. Posner generally uses the term catastrophe to mean “an event that is believed to have a very low probability of materializing but that if it does materialize will produce a harm so great and sudden as to seem discontinuous with the flow of events that preceded it” (Posner 2004, p. 6). Bostrom and Ćirković (2008, pp. 2-3) define global catastrophic risks as risks of serious events (e.g., with millions of fatalities or trillions of dollars of economic loss) with global scale.

Increasingly advanced and general AI models (including “foundation” AI models) such as GPT-3 could pose societal catastrophic risks, including potential for correlated robustness failures across multiple high-stakes application domains such as critical infrastructure; see, e.g., Bommasani (2021 pp. 115-116). As an example of near-term global catastrophic risks from AI, nuclear deterrence theorists have argued that developments in AI could increase the probability of nuclear war by reducing the stability of nuclear forces (Geist and Lohn 2018). Near-future AI systems will also permit designers to intervene at scales that have been previously possible only for larger human organizations like corporations and governments. This is now the focus of leading workshops on AI research; see, e.g., the *Political Economy of Reinforcement Learning Workshop*, *2021 Neural Information Processing Systems (NeurIPS) conference,* https://perls-

workshop.github.io/.

In the longer term, as advanced AI systems continue to grow in scale, generality, and capability, their potential to pose catastrophic risks could grow, e.g., because of potential for various kinds of misuse, or because the difficulty of safely controlling such systems would increase along with their capability. A number of leading computer scientists and AI researchers take such risks seriously, including DeepMind co-founder Shane Legg, and UC Berkeley professor Stuart Russell who co-authored the most-used textbook on AI; see, e.g., Legg (2008), Russell (2019), and Russell and Norvig (2020). The Asilomar AI Principles include several principles that address potential for global catastrophic risks from AI with growing capabilities, and have been signed by over 1700 AI and robotics researchers including Demis Hassabis, DeepMind Founder and CEO; Ilya Sutskever, OpenAI Co-Founder and Research Director; and Yann LeCun, Facebook Director of AI Research (FLI n.d.).

We also believe the first version of the RMF should consider future risks, including longer-term global catastrophic risks, because risk management decisions made in the near term could affect the long-term risks. The unique ability for the effects of AI systems to scale (as well as the increasing generality of their applicability, potential for use in high-stakes areas, etc.) means that if the RMF does not consider the long-term implications of AI from the beginning, it would miss a crucial portion of what is at stake in the coming years.

Our points on this cross-cutting topic relate to several specific topics in the RFI, including: challenges in risk management (Topic 1), definitions of AI characteristics such as safety (Topic 2), AI risk management principles (Topic 7), and risk to society (Topic 8).

References in this subsection:

Bender EM, Gebru T, McMillan-Major A, and Shmitchell S (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Conference on Fairness, Accountability, and Transparency*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA. https://doi.org/10.1145/3442188.3445922

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J,

Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P (2021), On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

Bostrom N, Ćirković MM, eds. (2008) Introduction. In *Global Catastrophic Risks*. Oxford University Press, Oxford, UK

Bradley C, Wingfield R, and Metzger M (2021) National Artificial Intelligence Strategies and Human Rights: A Review, Second Edition. Global Partners Digital and Stanford Global Digital Policy Incubator (April 2021): 1-70

Brkan M (2019). Artificial intelligence and democracy: The impact of disinformation, social bots and political targeting. *Delphi - Interdisciplinary Review of Emerging Technologies*, 2(2), 66–71. https://doi.org/10.21552/delphi/2019/2/4

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen GC, Steinhardt J, Flynn C, Ó hÉigeartaigh S, Beard S, Belfield H, Farquhar S, Lyle C, Crootof R, Evans O, Page M, Bryson J, Yampolskiy R, and Amodei D (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv*, https://arxiv.org/abs/1802.07228

Donahoe E and Metzger MM (2019) Artificial intelligence and human rights. *Journal of Democracy* 30 (2): 115-126

FLI (n.d.) Asilomar AI Principles. Future of Life Institute, https://futureoflife.org/ai-principles/

Geist E and Lohn AJ (2018) *How Might Artificial Intelligence Affect the Risk of Nuclear War?* RAND Corporation, https://www.rand.org/pubs/perspectives/PE296.html

Latonero M (2018) Governing Artificial Intelligence: Upholding Human Rights & Dignity. *Data & Society*, https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Legg S (2008) Machine Super Intelligence. Ph.D. Thesis. (University of Lugano, Switzerland.) http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

Mantelero A and Esposito S (2021) An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems. *Computer Law & Security Review* 41 https://doi.org/10.1016/j.clsr.2021.105561

Mozur P (2018). A Genocide Incited on Facebook, With Posts from Myanmar's Military. *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Nonnecke B and Dawson P (forthcoming) Human Rights Implications of Algorithmic Impact

Assessments: Priority Recommendations to Guide Effective Development and Use. Harvard Carr Center for Human Rights Policy Discussion Paper Series

Posner RA (2004), *Catastrophe: Risk and Response*. Oxford University Press, New York, New York

Russell S (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking

Russell S and Norvig P (2020) *Artificial Intelligence: A Modern Approach, 4th Edition.* Pearson

Recommendations

We recommend that the meaning of societal scale issues be expanded to include: risks to democracy and security such as polarization, extremism, mis- and disinformation, and social manipulation; risks to human rights and wellbeing including equity, environmental, and public health risks; and global catastrophic risks, including risks to large numbers of people caused by AI accidents, misuse, or unintended impacts in both the near- and long-term.

Risk assessment approaches focused on intended use cases have important limitations.

Comments

Consideration of intended AI use-cases is valuable and necessary, but not sufficient, for identification and assessment of important AI risks. We appreciate that NIST goes beyond focusing on intended use cases in the AI RMF RFI section Supplementary Information, Genesis for Development of the AI Risk Management Framework. That section states that “With broad and complex uses of AI, the Framework should consider risks from unintentional, unanticipated, or harmful outcomes that arise from intended uses, secondary uses, and misuses of the AI” and that the RMF should “be adaptable to many different organizations, AI technologies, lifecycle phases, sectors, and uses.” However, NIST does not clearly indicate scope beyond intended use cases when the NIST AI RMF RFI section Supplementary Information, AI RMF Development and Attributes, attribute 5, states that “...The Framework should assist those designing, developing, using, and evaluating AI to better manage AI risks for their intended use cases or scenarios.”

A focus on intended use cases could miss other foreseeable use cases and misuses. The limitations of a use case focused approach become more important as new AI systems become increasingly general in capability, with greater potential for adaptation to new uses (and misuses) across application domains. As an example of new AI systems with increasing generality of applicability, GPT-3 generated text with performance comparable to, or in some cases better than, task-specific fine-tuned systems (Brown et al. 2020). For discussion of the importance of considering potential misuse of AI, see, e.g., Brundage et al. (2018). The EU AI Act also includes the general idea of considering “reasonably foreseeable misuse” along with an

“intended purpose” of an AI system (EU 2021).

We recommend that the RMF include clear, usable guidance on identifying and assessing risks of potential uses, yielding risk management strategies that would be robust in the face of high uncertainty about future potential uses and misuses beyond the AI designers’ originally intended/planned uses. For example, to anticipate potential misuses, NIST should consider the cybersecurity concept of identifying and assessing risks of "misuse cases" or “abuse cases”, i.e., ways that either an adversary or authorized user could maliciously or accidentally misuse an information system, in addition to considering intended use cases for authorized users of an information system. Although no approach to identifying potential risks will do a perfect job of identifying all risks, we believe it will be worthwhile for the RMF to provide useful guidance on assessing such risks.

Microsoft provides some guidance on identifying potential types of harm, e.g., from intended uses, unintended uses, system errors, or misuses, as part of AI harms modeling; see Microsoft (2020). OpenAI has also codified the general idea of identifying abuse/misuse cases in their own AI safety best practices; their best practice #2 is "think like an adversary": see OpenAI (2020). OpenAI exemplified this approach in their 2019 announcement of GPT-3, which included several categories of potential misuse cases: see the “Policy Implications” section of OpenAI (2019).

More detailed guidance may not be available yet for identifying abuse/misuse cases for AI systems, nor for identification of broader sets of potential uses or secondary uses beyond the originally-envisioned intended uses. These may be gaps that could be addressed by the RMF, though it may require additional research and development. As an example of the level of detail of guidance that NIST should consider aiming for, see the documentation from the Open Web Application Security Project (OWASP) on identifying and prioritizing abuse cases for web-application software development (OWASP 2021). Recent NIST work such as NIST (2019) could be useful in defining abuse/misuse cases to consider for machine learning AI systems.

Our points on this cross-cutting topic relate to several specific RFI topics, including #1, 10, and 12, as well as goal #3.

References in this subsection:

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss, A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, and Amodei D (2020) Language Models are Few-Shot Learners. *arXiv*, https://arxiv.org/abs/2005.14165

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen GC, Steinhardt J, Flynn C, Ó hÉigeartaigh S, Beard S, Belfield H, Farquhar S, Lyle C, Crootof R, Evans O, Page M, Bryson J, Yampolskiy R, and

Amodei D (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv*, https://arxiv.org/abs/1802.07228

EU (2021) Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Microsoft (2020) Foundations of assessing harm. Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

NIST (2019) A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269, https://doi.org/10.6028/NIST.IR.8269-draft

OpenAI (2019) Better Language Models and Their Implications. OpenAI, https://openai.com/blog/better-language-models/

OpenAI (2020) Safety best practices. OpenAI, https://beta.openai.com/docs/safety-best-practices

OWASP (2021) Abuse Case Cheat Sheet. OWASP, https://cheatsheetseries.owasp.org/cheatsheets/Abuse_Case_Cheat_Sheet.html

Recommendations

We recommend that the RMF include clear, usable guidance on identifying and assessing risks of AI, yielding risk management strategies that would be robust despite high uncertainty about future potential uses and misuses beyond the AI designers' originally intended/planned uses.

The nascent but growing field of AI safety is providing insights about AI risks and risk management.

Comments

While much of the work in the field of AI safety is at an early stage, it has already yielded some general principles and tools that we expect could be useful to NIST stakeholders. For examples of resources that include concepts or tools for technical specialists in testing key aspects of AI safety, see Amodei et al. (2016), Ray et al. (2019), and OpenAI (2019a, 2019b).

Work adjacent to the field of AI safety has also highlighted the distinctive risks of formal models and real-world systems. This includes distinguishing the optimization of some represented task as part of a model vs. establishing control and stability over the dynamics of the domain in interaction with a given AI system. For a sociotechnical presentation that highlights important dimensions of this problem, see Andrus et al. (2020) and Dean et al. (2021).

The lack of clear or agreed-upon definitions for terms like "trustworthiness" and "safety" is now being examined by safety researchers (Dobbe et al. 2021). In addition, the Georgetown University Center for Security and Emerging Technology (CSET) briefs on AI safety provide summaries for broad audiences; see Rudner and Toner (2021a, 2021b, 2021c).

Our points on this cross-cutting topic relate to several specific topics in the RFI, including: challenges in risk management (Topic 1), definitions of AI characteristics such as safety (Topic 2), AI risk management principles (Topic 7), and risk to society (Topic 8).

References in this subsection:

Andrus M, Dean S, Gilbert TK, Lambert N, and Zick T (2020) AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks. *2020 IEEE International Symposium on Technology and Society* (ISTAS), pp. 72-79, doi: 10.1109/ISTAS50296.2020.9462193

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D (2016) Concrete Problems in AI Safety. *arXiv,* https://arxiv.org/abs/1606.06565

Dean S, Gilbert TK, Lambert N, and Zick T (2021), Axes for Sociotechnical Inquiry in AI Research. *IEEE Transactions on Technology and Society* 2(2), pp. 62-70, June 2021, doi: 10.1109/TTS.2021.3074097

Dobbe R, Gilbert TK, and Mintz Y (2021), Hard choices in artificial intelligence. *Artificial Intelligence* 300, https://doi.org/10.1016/j.artint.2021.103555

OpenAI (2019a) Safety Gym. OpenAI, https://openai.com/blog/safety-gym/

OpenAI (2019b) safety-gym. OpenAI, https://github.com/openai/safety-gym

Ray A, Achiam J, and Amodei D (2019) Benchmarking Safe Exploration in Deep Reinforcement Learning. OpenAI, https://cdn.openai.com/safexp-short.pdf

Rudner TGJ and Toner H (2021a) Key Concepts in AI Safety: An Overview. CSET, https://cset.georgetown.edu/wp-content/uploads/CSET-Key-Concepts-in-AI-Safety-An-Overview.pdf

Rudner TGJ and Toner H (2021b) Key Concepts in AI Safety: Robustness and Adversarial Examples. CSET, https://cset.georgetown.edu/wp-content/uploads/CSET-Key-Concepts-in-AI-Safety-Robustness-and-Adversarial-Examples.pdf

Rudner TGJ and Toner H (2021c) Key Concepts in AI Safety: Interpretability in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/CSET-Key-Concepts-in-AI-Safety-Interpretability-in-Machine-Learning.pdf

Recommendations

We recommend that the NIST Framework consider the nascent but growing field of AI safety in informing its deliberations.

NIST should continue to maintain awareness of progress in AI safety and other key fields, and update corresponding components of the RMF as needed.

Comments

The AI field has changed significantly over the last five years, and is likely to continue to change, perhaps even more dramatically. Ongoing research, particularly in such critical domains as AI safety, security, and capabilities will demand that the Framework is flexible enough to withstand potential shifts, and that NIST update corresponding components of the Framework as needed. To follow shifts across these fields and potential impact on the RMF, we recommend that NIST maintain close relationships with researchers in key fields, such as AI safety, security, and capabilities. These include researchers at three UC Berkeley research centers: the Center for Human-Compatible AI (CHAI), the Center for Long-Term Cybersecurity (CLTC), and the Center for Information Technology Research in the Interest of Society (CITRIS).

Our points on this cross-cutting topic relate to several specific topics in the RFI, including: challenges in risk management (Topic 1), definitions of AI characteristics such as safety (Topic 2), AI risk management methodologies (Topic 5), and risk to society (Topic 8).

Recommendations

We recommend that NIST maintain close relationships with researchers in key fields (including AI safety, security and capabilities) to follow shifts across these fields and potential impact on the RMF, and that NIST update corresponding components of the Framework as needed.

Coordination of standards for risk identification and mitigation, to the extent possible.

Comments

The NDAA requests that NIST ensure the Framework “align(s) with international standards, as appropriate.” Development and deployment of AI systems is often global. To better support efficiency and effectiveness in implementation of standards to identify and mitigate risks of AI, NIST should coordinate development of any AI standards with standards development organizations, including the Institute of Electrical and Electronics Engineers (IEEE), the International Standards Organization (ISO), the International Electrotechnical Commission

(IEC), the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), among others.

While standards may provide guidance on appropriate criteria to evaluate AI, it is important that standards are carefully developed to ensure relevant criteria are considered. If criteria in the Framework and corresponding standards are too narrow, they may inadvertently overlook potential risks. NIST's commitment to a flexible Framework that is consistently updated is critical to ensure appropriate identification and mitigation of risks.

Our points on this cross-cutting topic relate to AI RMF attribute #7, as well as RFI topics #1 and #5.

Recommendations

We recommend that NIST be explicit about how and where the RMF will incorporate and coordinate with existing and future AI standards development and risk assessment.

Specific RMF RFI Topics

RFI Topic 1 (The Greatest Challenges in Improving how AI Actors Manage AI-related Risks)

Comments

A general challenge is the identification, assessment and prioritization of risks that could have high consequences for society but may seem to be outside the typical scope of consideration by an organization's AI designers. One reason is that many high-consequence risks would involve novel or low-probability events, or systemic risks, that could seem very unlikely or outside the scope of the organization's direct responsibility. Moreover, organizations have limited resources for risk identification and risk mitigation. Furthermore, guidance available on identifying and assessing low-probability, high-consequence risks is likely less standardized and straightforward than typical guidance for identifying and assessing more common types of events (e.g., for standard information-system risk assessment). Thus, the RMF presents an opportunity for NIST to address these gaps and to guide organizations to consider risks of events with high consequences for society. The RMF also represents an opportunity within a voluntary framework to remind organizations of reasons why they should consider events with impacts to society, e.g., identifying risks to the organization's reputation if an AI project becomes associated with undesirable societal-level outcomes.

However, there are substantial challenges in addressing risks to society within a voluntary framework. Yeung (2021, p. 20) argues that such approaches as taken in the voluntary Privacy Framework may not be sufficient for the AI RMF: "Because [risks from use of AI systems] might cause physical harm or violate fundamental values, NIST should also incorporate more stringent

elements in the AI risk management framework than were in the privacy framework." As one way to address such challenges with voluntary frameworks, we suggest NIST consider coordinating guidance and other policy instruments including standards, at least for some domains. This could include collective proprietary attention to known risks, structured audits to help monitor poorly-understood domain dynamics, and/or certifications preceding deployment in high-risk settings.

Another challenge is uncertainty associated with the speed and degree of changes in AI capabilities. AI systems continue to become increasingly advanced, powerful, and impactful, sometimes much sooner than most AI researchers expect. It would be valuable for the RMF to include guidance for future-proofing of risk management, e.g., by looking further over time horizons than typical, and to consider potential for events such as reaching AI capability milestones that are not expected until further into the future.

An additional challenge is that managing AI risk is a distinct problem space from minimizing model bias, as models comprise just one end of a highly complex organizational "stack" or workflow (Andrus et al. 2020).

Risk communication that goes beyond narrow notification requirements also poses significant challenges to digital platforms, but there are some key lessons from other sectors and the scientific literature, and best practices that have emerged. See a preliminary roadmap in the CLTC report, Newman et al. (2020).

References in this subsection:

Andrus M, Dean S, Gilbert TK, Lambert N, and Zick T (2020) AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks. *2020 IEEE International Symposium on Technology and Society* (ISTAS), pp. 72-79, doi: 10.1109/ISTAS50296.2020.9462193

Newman J, Cleveland A, Gordon G, and Weber S (2020), Designing Risk Communications: A Roadmap for Digital Platforms. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2020/12/Designing-Risk-Communications.pdf

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2021/08/AI_Risk_Impact_Assessments.pdf

Recommendations

We recommend that the RMF provide guidance on risk identification, assessment and prioritization processes to include risks that could have high consequences for society but may seem to AI designers to be outside the typical scope of consideration for their organization, such as events that would be novel or low-probability events, or systemic risks, or expected to be outside their typical time horizon.

RFI Topic 2 (Characteristics of AI Trustworthiness)

Comments

For definitions of AI safety (as well as reliability, robustness, security, and harmful outcomes from misuse), see AI safety research agendas and publications such as Amodei et al. (2016).

Part of the work of safety is to build systems that remain under human control and are demonstrably subject to human oversight and periodic external evaluation. For one prominent example of technical work in this direction, see Hadfield-Menell et al. (2016).

We suggest that NIST consider “assessment of generality” (i.e., assessment of the breadth of AI applicability/adaptability) as another important characteristic affecting trustworthiness of an AI system, or perhaps as a factor affecting one or more of the AI trustworthiness characteristics NIST has already outlined. If an AI has not undergone any assessment of its generality, that would suggest lower trustworthiness. If assessment indicates high generality of an AI, we expect it would be appropriate to conduct more in-depth risk assessment, more assessment of use cases beyond the originally intended use cases, longer time horizons in risk assessment, more continuing assessment, etc. (Ideally, a generality assessment process would be quick and low-cost for the majority of AI with low generality, while accurately identifying the smaller number of AI with high generality.) For discussion of AI generality, see e.g. Bommasani et al. (2021).

For definitions of explainability, it is important to understand how the term has been used differently by various stakeholders and how in practice it has often failed to meet its objectives (Newman 2021). The definition of fairness is similarly contested (Mulligan et al. 2019).

In the definition of terms such as explainability, it is critical to consult with a diverse set of stakeholders to account for diverging uses of terms. On the topic of stakeholder engagement, one proposal to manage risks more effectively, reliably, and safely (Dobbe et al 2021) is to incorporate feedback from stakeholders, including risk-sensitive groups, democratizing the structure of AI pipelines. Dobbe et al (2021) provides a sociotechnical lexicon of terms and relevant dilemmas throughout AI development, as well as analysis of vagueness in AI development and how stakeholder input is needed to resolve it appropriately.

References in this subsection:

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D (2016) Concrete Problems in AI Safety. *arXiv,* https://arxiv.org/abs/1606.06565

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J,

Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P (2021) On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

Dobbe R, Gilbert TK, and Mintz Y (2021), Hard choices in artificial intelligence. *Artificial Intelligence* 300, https://doi.org/10.1016/j.artint.2021.103555

Hadfield-Menell D, Dragan A, Abbeel P, and Russell S (2016) Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29: 3909-3917, https://papers.nips.cc/paper/2016/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf

Mulligan DK, Kroll JA, Kohli N, and Wong RY (2019) This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 119 (November 2019), https://doi.org/10.1145/335922

Newman J (2021) Explainability won't save AI, *Brookings TechStream*, https://www.brookings.edu/techstream/explainability-wont-save-ai/

Recommendations

We recommend that NIST consult with a diverse set of stakeholders, including risk-sensitive groups, for input such as on definitions of key terms to better understand how the terms have been used differently by various stakeholders.

We also recommend that NIST consider "assessment of generality" (i.e. assessment of the breadth of AI applicability/adaptability) as another important characteristic affecting trustworthiness of an AI, or perhaps as a factor affecting one or more of the AI trustworthiness characteristics NIST has already outlined.

RFI Topic 3 (AI Principles)

Comments

Additional principles which should be considered are sustainability and inclusivity. For example, one of the OECD AI principles is, "AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being." Other AI risk and impact frameworks have

also included these considerations (Yeung 2021).

Over 170 sets of ethical AI guidelines have been developed (Algorithmwatch.org 2020). A growing consensus is emerging around the following principles: accountability, privacy and security, transparency and explainability, fairness and non-discrimination, professional responsibility, human control, and the promotion of human values such as civil and human rights.

Organizations are taking concrete steps to operationalize AI principles. For example, the OECD Network of Experts on AI is creating a database of tools and practices to implement the OECD AI Principles (OECD 2021). For a more in depth case study on how organizations such as Microsoft are defining and managing AI principles, see Newman (2020).

Finally, we recommend that NIST clarify two items in the RMF RFI regarding NIST's use of the terms "characteristics" and "principles". First, we recommend that the difference between principles and characteristics is made more clear. Second, where the RFI states that "These characteristics and principles are generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services", we recommend you clarify to what extent NIST meant "considered by the public", or "considered by experts", or both; differentiating expert and public evaluations of trustworthiness seems both descriptively salient and normatively appropriate. (This relates to RFI section Supplementary Information: Genesis for Development of the AI Risk Management Framework.)

References in this subsection:

Algorithmwatch.org (2020) AI Ethics Guidelines Global Inventory. Algorithmwatch.org, https://inventory.algorithmwatch.org/

Newman J (2020) Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles. CLTC, https://cltc.berkeley.edu/ai-decision-points/

OECD (2021), Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems. *OECD Digital Economy Papers*, No. 312, OECD Publishing, Paris, https://doi.org/10.1787/008232ec-en

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2021/08/AI_Risk_Impact_Assessments.pdf

Recommendations

We recommend that NIST consider including principles of sustainability and inclusivity. We also recommend that NIST clarify two items in the RMF RFI regarding NIST's use of the terms "characteristics" and "principles": 1. That the difference between principles and characteristics is made more clear, and 2. Where the RFI states that "These characteristics and principles are

generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services", we recommend you clarify to what extent NIST meant "considered by the public", or "considered by experts", or both.

RFI Topic 4 (AI Risks in Organizations' Enterprise Risk Management)

Comments

Research on organizational safety standards and the incorporation of AI technologies into the commercial aviation industry reveals how the opaque, unpredictable, and accident-prone nature of AI technologies results in slow adoption in safety critical domains. There is demand for collaborative AI safety standards that meet rather than relax aviation's high safety standards (Hunt 2020).

References in this subsection:

Hunt W (2020) The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf

RFI Topic 5 (Standards, Frameworks, Models, Methodologies, Tools, Guidelines and Best Practices)

Comments

For effective risk identification, one best practice is to have risk identification processes performed by a team that is diverse, multidisciplinary, representing multiple departments of the organization, as well as including a correspondingly diverse set of stakeholders from outside the organization. See, e.g., guidance on including stakeholders during project risk identification (PMI 2017, section 11.2), as well as guidance on the ranges of types of stakeholders to include when identifying potential types of AI harm (Microsoft 2020). As we mentioned previously, one proposal to manage risks more effectively, reliably, and safely is to incorporate feedback from stakeholders and risk-sensitive groups, democratizing the structure of AI pipelines (Dobbe et al. 2021). The diversity of perspectives from such approaches can help identify a greater breadth and depth of risks that otherwise could be missed by a team without the same perspectives.

It would be valuable for the Framework to include templates and definitions to facilitate information sharing on AI risk factors and incidents. Standardized tools for sharing information about incidents and risk factors could reduce costs and increase value of efforts to identify, assess, prioritize, mitigate, and communicate AI risk. For AI incident reporting, one leading effort is the Partnership on AI's AI Incident Database (AIID n.d). Reporting on AI risk factors potentially could adapt procedures and templates currently used in the cybersecurity community for vulnerability disclosure. NIST could provide standardized reporting formats or other means to

help AI developers share information in consistently beneficial ways.

As mandated in the NDAA, NIST should align its efforts with international standards, as applicable. In doing so, NIST will support the development of standards that support greater efficiency and effectiveness in risk mitigation. We recommend that NIST review the work of the IEEE Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), which is developing specifications including metrics and processes for addressing transparency, accountability, and algorithmic bias in autonomous and intelligent systems (ECPAIS 2021). We also recommend NIST review the work of the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) joint committee on artificial intelligence (ISO/IEC JTC1/SC 42), which is developing a conformity assessment standard for AI risk management (ISO 2021). The European Commission issued a report in 2021 that outlines relevant standards from the IEEE, ISO, and other standards development organizations that support compliance with principles outlined in the EU AI Act, including standards for appropriate data governance; risk management; technical data and record keeping; transparency and accountability; human oversight; accuracy, robustness, and cybersecurity (Nativi and De Nigris 2021). Yeung (2021) compares AI risk and impact assessment approaches of the EU AI Act and Canada's Directive on Automated Decision-Making, as well as of New Zealand, Germany, and San Francisco.

There is also a need for accountability tools and metrics that are suited to the risks of actual existing systems that have already been (or are likely to be) deployed—including the API, licenses, and data usage—in addition to and beyond the potential for statistical bias in formal models. Papers that speak to this perspective and acknowledge the distinction between systems and models include: Mitchell et al. (2019), Raji et al. (2020), and Paullada et al. (2020).

References in this subsection:

AIID (n.d.) AI Incident Database, Partnership on AI, https://incidentdatabase.ai/

Dobbe R, Gilbert TK, and Mintz Y (2021) Hard choices in artificial intelligence. *Artificial Intelligence* 300, https://doi.org/10.1016/j.artint.2021.103555

ECPAIS (2021) IEEE,
https://standards.ieee.org/industry-connections/ecpais.html

ISO (2021) ISO/IEC CD 23894.2 Information Technology — Artificial Intelligence — Risk Management, https://www.iso.org/standard/77304.html

Microsoft (2020) Foundations of assessing harm, Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model Cards for Model Reporting, in *Conference on Fairness, Accountability, and*

Transparency 2019, January 29–31, 2019, Atlanta, GA, USA. ACM, https://doi.org/10.1145/3287560

Nativi S and De Nigris S (2021) AI Watch: AI Standardisation Landscape state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40325-8, doi:10.2760/376602, JRC125952

Paullada A, Raji ID, Bender EM, Denton E, and Hanna A (2020) Data and its (dis)contents: A survey of dataset development and use in machine learning research. *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses*, Virtual, https://arxiv.org/pdf/2012.05345.pdf

PMI (2017) Guide to the Project Management Body of Knowledge, Sixth Edition, Project Management Institute, Newtown Square, PA

Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, and Barnes P (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. in *Conference on Fairness, Accountability, and Transparency 2020*, January 27–30, 2020, Barcelona, Spain, ACM, https://doi.org/10.1145/3351095

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2021/08/AI_Risk_Impact_Assessments.pdf

Recommendations

We recommend that NIST consider having the RMF include guidance to have risk identification processes performed by a team that is diverse, multidisciplinary, representing multiple departments of the organization, as well as including a correspondingly diverse set of stakeholders from outside the organization.

We also recommend that the RMF include standardized templates for reporting information on AI risk factors and incidents, that AI developers could adopt voluntarily.

RFI Topic 7 (Alignment with Other Efforts)

Comments

For a comparative analysis of AI risk and impact assessments from five regions around the world including Canada, New Zealand, Germany, the European Union, and San Francisco, California, see Yeung (2021).

Please also see our discussion above of standards related to NIST AI RMF RFI topic #5.

References in this subsection:

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments, CLTC, https://cltc.berkeley.edu/2021/08/09/guidance-for-the-development-of-ai-risk-and-impact-assessments/

RFI Topic 8 (Inclusiveness)

Comments

Case studies documented in Newman (2020) detail how institutions including Microsoft and OpenAI have tried to improve the inclusiveness of AI design, development, use, and evaluation and also reduce and manage the risk of potential negative impacts. At Microsoft for example, the Responsible AI Program includes the AETHER Committee, the Office of Responsible AI, a Responsible AI Standard, and a Responsible AI Champs community. Microsoft researchers have also documented the role of checklists in AI ethics and worked on “harms modeling” designed to help researchers anticipate the potential for harm and identify gaps in products that could put people at risk (Madaio et al. 2020, Microsoft 2020).

References in this subsection:

Newman J (2020) Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles, CLTC, https://cltc.berkeley.edu/ai-decision-points/

Madaio M et al. (2020) Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, https://dl.acm.org/doi/abs/10.1145/3313831.3376445

Microsoft (2020) Foundations of assessing harm, Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

RFI Topic 9 (Attributes for RMF)

Comments

While the RMF attributes list currently includes using plain language that is understandable by a broad audience, it does not explicitly include being user-friendly more broadly. Enabling ease of use for diverse stakeholders - for example by including implementation guides - is advised in order to help NIST achieve its goals for the AI RMF.

Recommendations

We recommend that NIST consider adding usability as an attribute of the AI RMF.

RFI Topic 10 (Structuring the Framework)

Comments

We commend NIST for planning to take an iterative approach with AI RMF development. We expect that appropriate, net-beneficial guidance addressing many key concepts (e.g., for some technical aspects of safety) may require more time to develop than would be feasible for inclusion in the initial Framework.

We suggest that NIST consider clarifying its planned procedures for making RMF updates (how often, under what conditions, decision criteria), and how it aims to balance flexibility with standard-setting authority.

For recommendations on linking the AI risk framework to procurement and purchasing decisions, see Yeung (2021).

Yeung (2021, p.19) also discusses how the NIST Privacy Framework, as a voluntary framework, reminds organizations of reasons and incentives to consider risks affecting external stakeholders: "the framework points out how privacy risks can ... impact the organization, such as its reputation taking a hit or revenue loss from customers moving elsewhere. This linkage to organizational impact helps to provide parity between privacy risks and other risks that organizations are managing and leads to more informed decision-making." Similarly, the NIST Cybersecurity Framework also mentions that cybersecurity incidents can affect an organization's reputation. However, Yeung (2021, p. 20) also argues that such approaches taken in the Privacy Framework may not be sufficient for the AI RMF: "Because [risks from use of AI systems] might cause physical harm or violate fundamental values, NIST should also incorporate more stringent elements in the AI risk management framework than were in the privacy framework."

Analytic dimensions of AI risks and possible domain manifestations are now being explored and mapped by technical and sociotechnical researchers. See Dean et al. (2021).

References in this subsection:

Dean S, Gilbert TK, Lambert N and Zick T (2021) Axes for Sociotechnical Inquiry in AI Research, in *IEEE Transactions on Technology and Society* 2 (2), pp. 62-70, June 2021, doi: 10.1109/TTS.2021.3074097

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments, CLTC, https://cltc.berkeley.edu/2021/08/09/guidance-for-the-development-of-ai-risk-and-impact-assessments/

Recommendations

We recommend that NIST consider clarifying its planned procedures for making RMF updates (how often, under what conditions, decision criteria), and how it aims to balance flexibility with standard-setting authority.

RFI Topic 12 (Governance)

Comments

It would be very valuable for the Framework to include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. These should include guidance for who should be responsible for implementing the Framework within each organization, ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems, support for incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers, among other mechanisms. On auditing see, e.g., Raji et al. (2020); on AI incidents see the AI Incident Database (McGregor 2020) and Arnold and Toner (2021). We also recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety; related considerations are included in other AI risk frameworks (Yeung 2021).

For an example of a leading AI enterprise that reviews applications that would use their AI platform, and disallows unacceptable categories of use cases, see OpenAI (2020).

Assessment frameworks that address this include explorations of the problem of “trustworthy” mechanisms for verifying development claims and Z-inspection as a domain-specific approach to risk diagnostics. See Brundage et al. (2020) and Zicari et al. (2021).

We recommend that NIST include guidance on governance processes to support the successful implementation of the AI RMF. We recommend reviewing Moss et al. (2021), which outlines “10 constitutive components” of supporting accountability in impact assessments. NIST should provide guidance on ways to support accountability in the implementation of the RMF (e.g., suggesting personnel/management levels that will implement and oversee the RMF process).

We suggest NIST also consider providing guidance on the makeup of the design and development teams, e.g. according to the diagnostic, formalizer, rebuttal, and synecdoche roles (or at least on the relative importance of those roles in particular use cases) outlined in Abebe et al. (2020).

With AI systems growing increasingly complex, it becomes more difficult to assess whether an AI system constitutes safety risks or violations of human rights. At times, the system developers will be some of the only people in a position to assess the types and magnitudes of risks. Those developers should have options to raise concerns to outside authorities if organization-internal

channels seem insufficient, and whistleblowers should have appropriate protections.

References in this subsection:

Abebe R, Barocas S, Kleinberg J, Levy K, Raghavan M, and Robinson DG (2020) Roles for Computing in Social Change. In *Conference on Fairness, Accountability, and Transparency*, January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3351095.3372871

Arnold Z and Toner H (2021) AI Accidents: An Emerging Threat; What Could Happen and What to Do, CSET, https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, Ryffel T, Rubinovitz JB, Besiroglu T, Carugati F, Clark J, Eckersley P, de Haas S, Johnson M, Laurie B, Ingerman A, Krawczuk I, Askell A, Cammarota R, Lohn A, Krueger D, Stix C, Henderson P, Graham L, Prunkl C, Martin B, Seger E, Zilberman N, Ó hÉigeartaigh S, Kroeger F, Sastry G, Kagan R, Weller A, Tse B, Barnes E, Dafoe A, Scharre P, Herbert-Voss A, Rasser M, Sodhani S, Flynn C, Gilbert TK, Dyer L, Khan S, Bengio Y, and Anderljung M (2020) Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, *arXiv*, https://arxiv.org/abs/2004.07213

McGregor S (2020) When AI Systems Fail: Introducing the AI Incident Database, Partnership on AI,
https://partnershiponai.org/aiincidentdatabase/

Moss E, Watkins EA, Singh R, Elish MC, and Metcalf J (2021) Assembling Accountability: Algorithmic Impact Assessment For The Public Interest, Data & Society, https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf

OpenAI (2020) Use case guidelines, OpenAI, https://beta.openai.com/docs/use-case-guidelines

Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, and Barnes P (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, in *Conference on Fairness, Accountability, and Transparency 2020*, January 27–30, 2020, Barcelona, Spain, ACM, https://doi.org/10.1145/3351095

Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments, CLTC, https://cltc.berkeley.edu/wp-content/uploads/2021/08/AI_Risk_Impact_Assessments.pdf

Zicari RV, Brusseau J, Blomberg SN, Christensen HC, Coffee M, Ganapini MB, Gerke S, Gilbert TK, Hickman E, Hildt E, Holm S, Kühne U, Madai VI, Osika W, Spezzatti A, Schnebel E, Tithi JJ, Vetter D, Westerlund M, Wurth R, Amann J, Antun V, Beretta V, Bruneault F, Campano E,

Düdder B, Gallucci A, Goffi E, Haase CB, Hagendorff T, Kringen P, Möslein F, Ottenheimer D, Ozols M, Palazzani L, Petrin M, Tafur K, Tørresen J, Volland H, and Kararigas G (2021) On Assessing Trustworthy AI in Healthcare: Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Frontiers in Human Dynamics* (3) https://www.frontiersin.org/article/10.3389/fhumd.2021.673104

Recommendations

We strongly recommend that the Framework include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. These should include guidance for determining who should be responsible for implementing the Framework within each organization, ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems, support for incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers, among other mechanisms. We also recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.